基于生成对抗网络的智能音乐制作综述

马 丹，吴 跃

(电子科技大学，计算机学院，成都 610000)

摘 要：如何借助计算机算法进行音乐的自动或半自动化生成工作一直是人工智能领域的一个研究热点。近年来，随着深度学习技术的深入发展，使用基于神经网络并契合乐理先验知识的方法来生成高质量、多样性智能音乐的任务也引起了研究者的重视。其中，引入生成对抗机制以提升生成效果的工作取得了一定成果，同时也具备极大的提升空间。为了更好地推进后续研究工作，对相关领域的现有成果进行全面而系统的梳理、分析、总结具有比较重要的意义。首先对机器作曲的发展过程进行了回顾，对音乐领域常用的 GANs 相关重要模型进行了简要归纳介绍，对引入了生成对抗训练机制的音乐生成方法进行了重点分析，最后对该领域的现状进行了总结并进一步展望了未来的发展方向。

关键词：生成对抗网络；智能音乐；强化学习；模式塌陷

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2020.02.0030

Survey of intelligent music creation based on GANs

Ma Dan, Wu Yue

(*School of computer science & engineering, University of Electronic Science & Technology of China, Chengdu Sichuan* 610000, *China*)

Abstract: Recently, researchers pay more and more attentions for automatic or semi-automatic music generation based on computer algorithm. With the further development of deep learning, researchers start to focus on producing high-quality and multifarious-style music with neural networks and priori knowledge of music theory. Furthermore, several works introduced Generative Adversarial Networks (GANs for short) to try to improve the quality of the results. To summarize the important results in this area is meaningful and with guiding significance for the following works. The paper firstly reviewed the history of intelligent music, then listed related GANs model that are commonly applied in music creation, after that the paper analyzed some important works in this area. Finally, some observations were presented, and future work direction were prospected.

Key words: generative adversarial networks; intelligent music; reinforcement learning; mode collapse

0 相关工作

音乐作为艺术领域里一种重要的表达方式，体现了一系列人类所特有的思维模式，是规则性和创造性的统一结合体。一方面，音乐的构成天生基于一定的乐理规则，如旋律、节奏、调式、和弦、和声、复调、曲式等，不能满足乐理规则约束的乐曲往往在听觉悦耳度上欠佳且不能被大众接受。另一方面，单纯满足乐理约束的音乐不一定算是好音乐，音乐本身还承载了情感表达载体的重任，这就需要创作者不能墨守成规地进行规则堆砌，而需要在音乐中揉入创新性，使得生成的旋律不至于千篇一律，模式固定。

与此同时，通过计算机算法来自动生成音乐一直是人们较为关注的领域，对该领域的研究目的在于：

a)借助计算机算法来进行音乐创作可以降低制作门槛，节省人力及时间成本，一定程度上规避版权问题，并根据场景需求快速进行大量音乐制作，如影视剧中需要大量定制渲染剧情情感的旋律等。

b)由于音乐等艺术创作领域具备的规则性与创造性等特点，对智能音乐创作的研究可以很好地衡量和测试人工智能能力的性能。不少音乐生成领域相关工作在描述自身的模型算法实验结果时，均采用了组织自愿者进行听觉识别的检验方式，从真实性、悦耳性、创造性、趣味性等多方面进行考察统计，如文献[1]通过随机寻找 144 位测试者检验其 MuseGAN 模型生成音乐样本的质量。类似地，一些其他的算法作曲的工作[2~5]也采用了这种人力的评判办法。

一般来说，智能音乐的制作过程需要让人工干预的工作量最小化(minimal human intervention)[6]，该制作方式可以全自动(total automation)或半自动(partial automation)生成音乐。在输出的结果上既要满足基本的乐理先验知识，也要具备一定的算法创造性。文献[7]中，作者论述该过程是从计算模型中自主地制作连续音频信号或者离散的符号序列，而这些信号和序列必须满足乐理架构[8]。

早在上世纪 50 年代，人工智能技术刚处于萌芽时期，虽然受到数据和硬件性能等多方面的限制，人们也开始在智能作曲领域进行探索，并取得了一定的成果。早期主要以两种方式来生成智能音乐：

一是基于统计分析的方式，结合马尔可夫链等模型进行创作，如文献[9]最早使用大型计算机 Illiac 创作弦乐四重奏组曲，成为历史上第一个完全由计算机生成的音乐作品，作者使用马尔可夫链模型来产生有限控制的随机音符，并结合和声与复调的规则测试这些音符，对通过测试的‘元素材’进行修改合成传统音乐记谱的弦乐四重奏。

二是基于乐理规则做简单的模式匹配和机器学习，如文献[10]将乐理融入到机器学习中以生产音符。

近年来，随着深度学习及神经网络技术的不断深入发展，利用深度神经网络来生成音乐成为一个重要的研究方向

收稿日期：2020-02-15；**修回日期：**2020-04-04

作者简介：马丹(1980-)，男，四川成都人，博士研究生，主要研究方向为生成对抗网络、智能音乐生成(182448161@qq.com)；吴跃(1958-)，男，上海嘉定人，教授，博导，硕士，主要研究方向为网络计算.

[11~14]。深度学习通过学习训练集中的样本数据，借助神经元处理非线性拟合操作，对数据进行生成或判定。文献[14, 7]首次将音符时序性考虑在内，采用循环网络(recurrent networks)生成数据。文献[15]将音乐元素拆解为各种模式(mode)和浓度系数(density)，采用反向传播算法训练神经网络来拟合符合用户品味的音乐。文献[16]更进一步提出端到端的深度神经网络音乐生成方式。此后一些深度学习模型如 WaveNet[4]，MidiNet[5]，MuseGAN[1]等在这些研究成果的基础上，进一步提升了智能生成质量。

同时，一些与自然语言处理相关的技术也同时被引入到了音乐生成的领域[17, 18]。国内也有不少研究者将自然语言技术的扩展作为音乐生成的重要手段：文献[46]将字符级的循环网络结构用于音乐生成中，摆脱了传统的特征工程，实现了端到端的音乐输出；文献[47]基于栈式自编码器提取音符的隐式特征，并送入循环网络以生成音乐；单纯的声音信号或者自然语言处理与智能音乐处理相比有众多相似之处，构成音乐的最简单元素音符也具有上下文相关性，时序性等自然语言处理(NLP)相关的自然属性，不同之处在于：

a) 音乐存在多轨的概念，如吉他，钢琴，人声，和弦，贝斯等，而 NLP 处理不涉及多轨；

b) 音乐的走向受和弦、琶音、旋律、复音等乐理规则的控制[1]，所以音符的概率分布空间有其特殊的，有别于自然语言的先验性知识。

随着 AI 技术的不断进化更新，音乐创作的形式化技术也在其中得到了进一步的发展。近年来，生成对抗网络(generative adversarial networks，GANs)[19]在数据生成领域中收到越来越多的重视，依靠其强大的拟合能力和简单的训练推导过程(只需要做反向传播，避开了马尔可夫链反复采样)，被引入到了众多应用领域(特别是计算机视觉)。GANs 在图形生成[51]，图形压缩[50]，语音生成[4]，超分辨率还原[52]的场景中得到了广泛应用。文献[20]中，作者指出：“GAN 也能生成文本，可以进行对话生成、机器翻译、语音生成等. 同时，GAN 在其他领域也有涉及，比如生成音乐、密码破译等. 但是 GAN 在其他领域的应用效果并不显著，那么，如何提高 GAN 在其他领域的应用效果将值得深入研究，使生成对抗网络在人工智能方面大放异彩.”。利用 GANs 来进行语音合成[4][49]及音乐创作的研究也一直处于探索发展的阶段，并取得了不少成果。通过调研发现，对算法作曲的综述文献数量相对较多[8, 21, 22]，然而对基于深度学习特别是基于生成对抗网络的方式来生成音乐的综述文献相对缺失。截止到论文撰写时，没有相关综述对以下几点进行过全面考量和总结：a) 对近年来利用 GANs 网络进行音乐生成领域的技术进展进行介绍；b) 对相关研究成果的模型结构、数据形式、训练技巧，效果及缺陷等进行全面总结；c) 对相关发展趋势进行展望。本文重点关注上述几点，力求为后续研究者提供研究依据。

1 GANs 及其在音乐生成领域的重要衍生模型

2014 年，文献[19]首次提出了生成对抗网络的概念，用于从训练样本中学习出新样本。该网络由两个子网络组成，分别是生成器(generator)和判别器(discriminator)。生成器 G 的任务是让模型尽可能拟合真实的训练数据分布，而判别器 D 致力于区分输入的数据是来源于真实的数据还是由生成器制造的假数据。两者在训练过程中，不断提升自身能力，最终达到一种纳什均衡的理想状态，使生成器生成的数据最大可能地贴近真实数据分布。GANs 的基本损失函数如下：

$$\min_G \max_D V(D,G) = \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{1}$$

GANs 的对抗训练与其说是模型不如说是提供了一种框架和思想，理论上来说，生成器和判别器可以用任意可微分的函数的实现,并不局限于多层感知机或卷积神经网络。

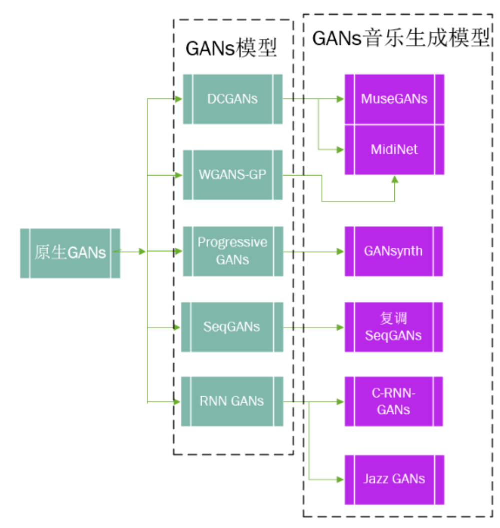


图 1 GANs 及其衍生的音乐生成模型

Fig. 1 Gans and compositional models derived from them

1.1 DCGANs

DCGANs 全称为深度卷积生成对抗网络(deep convolutional GANs)[23]。区别于原始 GANs 中用多层感知机实现的方式，DCGANs 的生成器 G 和判别器 D 均使用深度卷积网络来实现，并做了一些工程化改进技巧，如去除池化层，使用 Batch Normalization 协助模型快速收敛等，非常适合对多维张量数据的处理。DCGANs 在图形处理方面得到较为广泛的应用，同时，将音频数据进行矩阵化预处理后，也可以使用 DCGANs 来进行处理[5]。

1.2 Progressive GANs

NVIDIA 发布的 Progressive GANs[24]提出了一种动态增加模型训练层数的方法，提升模型生成质量。模型从低分辨率的图像(如 4*4)开始同时训练 G 和 D，随后在训练过程中动态地逐渐增加两者的层数。该模型在训练中首先聚焦图像的大致框架(低分辨率所覆盖的信息)，然后逐步将注意力转移到具体细节上去(高分辨率所蕴藏的信息)。整个添加过程对 G 和 D 来说是对称的，并且更新层数后保持已有层的参数继续训练，既保证了高分辨率合成像素的稳定性，也加快了训练的速度。文献[38]采用了该对抗训练方式来生成音乐旋律。

1.3 WGAN-GP

GANs 虽然在理论上对数据生成质量有极强自证性[19]，但是在实际训练中往往难以在收敛性和可靠性上达到较为理想的程度。文献[26]通过刨析生成器拟合原始数据的原理和 Loss 函数的本身缺陷入手，解释了 GANs 训练不稳定的原因。D 一旦被训练得太好，则 G 无法得到足够的梯度信息继续优化，甚至出现梯度消失的情况；而如果 D 的识别能力不足，同样不能让 G 进行有效的学习。同时文章也从衡量散度距离的角度出发指出了采用其他形式的 loss 函数训练 GANs 的不足，并解释了造成模式塌陷，训练难以收敛的原因。

文献[27]在此基础上提出采用 Earth-Mover(EM)距离，又称 Wasserstein 距离来衡量真实数据和生成数据的分布，并以拉近 EM 距离作为 GANs 的训练目标。作者采用满足 1-Lipschitz 限制的 loss 函数来替代求解 EM 距离。为满足 1-Lipschitz 限制，在实际网络训练中，模型的每次更新都要把参数截断在某个范围(即 weight clipping)。以此种方式训练的模型即 Wasserstein GANs，也即 WGANs。

WGANs 虽然解决了原始 GANs 的一些问题，但本身也

存在一些不足。后续文献[28]指出 weight clipping 的方案可能会导致大部分的模型参数都被集中设置为正负 0.01。为了解决该问题，作者又在原 WGANs 的基础上给 loss 函数加入梯度惩罚项(gradient penalty ，GP)，新的网络模型就叫 WGAN-GP，如公式所示。等式右边第一项即为原始的 WGANs 的距离衡量函数，本质上是用随机变量的数字特征(均值)的距离来表征两个分布的距离。等式第二项即为梯度惩罚项，保证了 G 的生成数据在向真实数据 x 靠近的过程中，D(G(z))不超过 D(x)，保持梯度的稳定性。文献[5]采用式(2)中基于 EM 距离的 loss 函数对 midi 音乐进行了生成训练，较好地提升了模型的稳定性。

$$L=(\underset{\tilde{x}\sim p_g}{E}[D(\tilde{x})]-\underset{x\sim p_r}{E}[D(x)])+\lambda\underset{\hat{x}\sim p_{\hat{x}}}{E}[(\|\nabla_{\hat{x}}D(\hat{x})\|_2-1)^2] \tag{2}$$

1.4　SeqGANs

长期以来 GANs 的学习对象都是以图像构建的连续张量数据，对于离散数据(如文字，音符)则因为梯度回传困难的原因进展缓慢。文献[29]结合了强化学习和 GANs 对抗训练的思想，开创了一个崭新的训练模式。作者把整个 GANs 网络看做一个强化学习系统，D(基于 CNN)输出得分作为强化学习中的奖励信号(reward signal)以识别数据的真伪，并采用 Policy Gradient 算法对 G(基于 RNN)进行更新，解决了离散数据训练过程中无法回传梯度的难题。同时，利用蒙特卡洛树搜索(Monte Carlo tree search，MCTS)的思想，对 G 生成的离散数据序列进行序列补全(roll out)，因此 D 就可以对任意时刻的非完整序列进行评估。解决了这两个问题后，SeqGANs 将对抗生成数据的问题统一到强化学习标准的 action-value 求解模型中，分别应用在文字和音符的生成示例中。

作者基于 SeqGANs 模型，对 Nottingham 数据集(http://www.iro.umontreal.ca/˜lisa/deep/data)中的 695 首 midi 歌曲样本进行预处理，提取音调独奏 solo 音轨进行训练，使用均方误差 MSE[30]作为测评量化标准，结果显示生成音符质量好于最大似然估计 MLE 所生成音符。SeqGANs 作为标准架构介绍，作者并未将应用生成作为重点铺开，示例中对音乐的生成也仅仅限制于单调音乐(monophonic melody)，后续工作[2]中进一步应用 SeqGANs 对音乐元素进行 word2vector 编码训练，生成复调音乐(polyphonic music)。

2　基于 GANs 的音乐生成方法

近年来，基于 GANs 的音乐生成方法数量虽然不太多，但很多重要方法值得借鉴，本章对相关方法进行了介绍，重点关注其算法，模型架构，数据表达，效果及缺陷等方面，并在表 1 中进行了相关对比。

表 1　基于 GANs 的重要音乐生成模型对比

Tab. 1　Comparison of various important compositional models based on GANS

模型	主干 GAN 架构	网络模型	效果	局限性
C-RNN-GAN	原生 GANs	RNN	相较于非对抗训练模型提升显著	听觉感受不佳
JazzGAN	原生 GANs	RNN	能生成质量较高的 Jazz 风格音乐，效果优于 C-RNN-GAN 及 seqGAN	生成音乐风格受限
MidiNet	DCGAN	CNN	效果优于 C-RNN-GAN	音符力度未参与训练，无法识别长拍音符和短拍连续按键音符的区别，表现力弱
MuseGAN	WGAN	CNN	测试发现三种模型中 Hybrid model 效果最优	没有将和弦的先验条件考虑在内
GANsynth	ProgressiveGANs WGANs_GP	CNN	相对于 WaveNet 生成速度极快，且质量上有较大的优势	GANs 的固有缺陷，如模式坍塌
SeqGAN	SeqGAN	RNN CNN	/	仅能生成单调音乐
复调 SeqGAN	SeqGAN	RNN CNN	/	模式坍塌，生成效率低

2.1　主干网络为 RNN 的方法

循环神经网络架构(recurrent neural networks ,RNNs)通常用于对序列数据建模，如自然语言处理中预测下一个单词，或在音乐生成中对下一个音符进行推断[31, 32]。

a) C-RNN-GAN

文献[33]提出了 C-RNN-GAN 的模型，作者受文献[34]中 coarse-to-fine 生成图片的启发，使用 RNN 网络(准确来说是双向长短时记忆网络 Bi-LSTM)来实现 GANs 中的生成器和判别器，生成符合 midi 标准的时序性连续数据。

C-RNN-GAN 采用了标准的原始 GANs 损失函数进行训练，由 G 生成连续化序列数据，D 区分生成数据和原始数据的真伪，如图 X 所示。作者构建了一个四元组数据，分别用以表征音符长度，音符频率，音符力度时长(tone lengths, frequencies, intensities, and timing)。

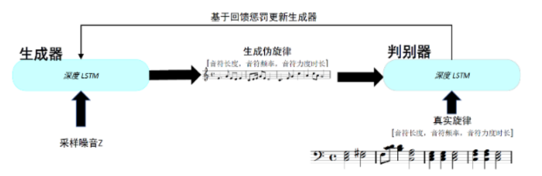


图 2　C-RNN-GAN 流程示意图

Fig. 2　Illustrion of C-RNN-GAN’s pipeline

值得注意的是，该模型在训练过程中使用了一些技巧，提升了训练质量，如 a)使用 L2 正则对 G 和 D 的权重做正则化约束；b) 在训练初期单独对 G 进行了 6 个 epoch 的训练，在该预训练过程中，对采样的序列长度做了管理，从小序列开始逐渐加大，最后变成长序列，最终提升了训练的稳定性；c) 采用了文献[35]中的冻结技巧，当 D 或 G 的能力对比达到阶段性不平衡时，可能会造成弱势一方的梯度消失，此时应对过于强大的一方实施冻结；d) 采用了文献[35]中的特征匹配技巧，将 G 的目标函数替换为使真假样本的特征差值最小化。

C-RNN-GAN 作为将对抗思想引入到音乐生成工作中较为早期的尝试，从人耳听觉感受上来说，其生成结果完全不能和真实样本相提并论。但是，作者将对抗机制取消，以单纯 RNN 网络生成的音符作为评测标准(即直接使用架构中的 G 来进行生成)，从几个维度来进行对比实验后，发觉生成的乐曲在多样性上有较为显著地提升，音程跨度更大，音调更为合理，复音更为丰富，更贴近于原始训练样本。

b) JazzGAN

文献[36]为了提供一个爵士音乐的教学工具，聚焦于利用生成对抗网络来生成爵士音乐，其主干模型采用的是循环神经网络 RNN，也是首个基于 RNN 和 GAN 使用离散化序

列(确定的音符种类)进行音乐生成的模型。相比于一般的音乐来说，爵士乐具有调性改变频繁，节奏不依常规，离弦音符(不在和弦范围内音符)较多的特性。以往模型所生成的传统音乐中缺乏爵士类音乐的上述创造性。该文在模型架构上没有做太大的更新，沿用了原生的 LSTM，但根据爵士乐自身特点提出了一系列更有针对性的训练数据封装方式及乐曲质量评定标准。

在数据方面，与 C-RNN-GAN 不同的是，JazzGAN 生成的是离散的音符种类(discrete pitch classes)，而不是回归浮点数类型的音符频率(real-valued frequencies)，将休止符当作一个特定种类解决前者无法处理休止符的问题。作者使用 RNN 搭配 GAN 架构测试了三种不同的旋律节奏编码方式，分别为：时间-步长编码方式(首次在 RNN 搭配 GAN 结构中使用该编码方式)，音符时长编码方式(C-RNN-GAN 和 SeqGAN[29]的编码方式)和音符节奏位置编码方式(和音符时长编码不同之处在于预测音符的结束时间值而不是音符时长值)。

在质量评定标准方面，JazzGAN 沿用了 C-RNN-GAN 的调性一致性(scale consistency)，重复音符计数(repetition counts)和音域跨度(tone spans)三个标准，也综合了 MuseGAN[1]提出的几个验证标准：a)单个序列中生成音符种类数量；b)合格的音符(在 MuseGAN 中被定义为时长超过三十二分之一音符的音符，在 JazzGAN 中被定义为时长超过 48 个时间步的音符)。

实验结果表明音符节奏位置编码方式在大多数性能标准的评测中都领先于 C-RNN-GAN 及 SeqGAN。

2.2　主干网络为 CNN 的方法

卷积神经网络(convolutional neural network, CNN)具备的参数共享性，特征平移不变性，邻近数据特征捕捉等性质使得其在图像张量数据的处理上具备了得天独厚的优势。同时，CNN 在训练速度及并行性上显著优于循环迭代架构的 RNN，但是由于 CNN 的卷积感受野有限，增大卷积核尺寸对模型整体效率影响显著，且 CNN 的数据组织受限，所以未能广泛应用在音乐制作上。2016 年，Deepmind 提出了空洞卷积(dilated convolution)的概念以提升卷积的感受野[4]，使模型不仅仅关注于数据的局部相关性，而对具备一定时间跨度的时序数据也具备特征提取的功能。 基于空洞卷积搭建的 WaveNet 模型[4]，给音频及音乐等序列数据的处理方式提供了新的思路。

1) MidiNet

受 WaveNet 的启发，文献[5]的作者尝试用 CNN 来生成音乐并在训练系统中加入对抗训练机制，该模型被称为 MidiNet。MidiNet 没有采用纯粹的连续时间序列生成旋律，而是以小节为单位(bar)，对小节进行逐一生成(one by one)，同时将前序生成的 bar 作为条件，输入到下个 G 的生成过程中。这样 MidiNet 模型既可以从 scratch(不带前置条件的噪声数据)中生成旋律，也能从前置音乐片段中生成旋律(需编码为 1D 或 2D 向量)。为了适应 CNN 的数据处理，作者将 midi 文件切分成 bar，并组织成 h*w 的矩阵，h 表示需要考察的 midi 音符的数量，w 是一个 bar 中的时间印戳单位(time_step)，多个轨道就需要组织多个矩阵数据。模型的主干网络实现采用 DCGANs[23]，同时在网络中加入了条件控制机制，使得一些先验知识(如前序旋律)可以被编码为 1D 或 2D 向量融合到网络的不同层中， 在 GANs 的训练过程中使用了文献[35]中提出的特征匹配和单边标签平滑的技巧，即在计算 GANs 标准损失函数的交叉熵的时候，正样本的标签使用 0.9 来替代 1，使梯度反传更加平滑，模型收敛性更强。

通过对 8 个小节的音乐生成测试，作者发现该模型生成的旋律对比常规 RNN 模型的生成结果来说，更真实悦耳，并且在趣味性上具备优势。但 MidiNet 没有有效将音符力度融入训练，也无法识别长拍音符和短拍连续按键音符的区别，在表现力上依然存在一定缺陷。

2) MuseGAN

文献[1]提出了音乐生成在时序性，多轨道生成和乐理复杂性上与一般的像素生成有显著区别，并提出了三个基于生成对抗训练的模型来生成音乐：Jamming 模型 Composer 模型 和 Hybrid 模型。每个模型的主干网络都采用了更易于训练的 WGAN[27]作为基准网络。Jamming model 中每个音轨都由独立的 G 来生成，并由独立的 D 来对抗识别；Composer model 中所有音轨统一由一个 G 来生成，也有统一的 D 来对抗识别；而 Hybrid model 在 Jamming model 的分立 G 生成基础上更进一步，每个音轨生成时附带有额外的输入信息，以保证整个多轨道旋律既有自身的特征也有全局统一调配的和谐性，实验效果也说明此种架构的生成结果效果更优。上述结构目的在于怎样在不同音轨中生成单个的小节 bar，但 bar 与 bar 之间的时序关联需要其他的结构来补充生成。

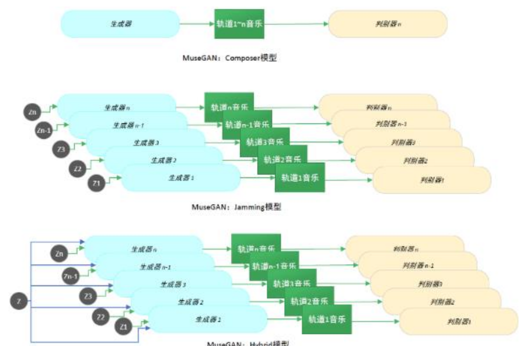


图 3　MuseGAN 的三种模型示意图

Fig. 3　Illustration of 3 types of musegan

与文献[5]一致的是，MuseGAN 也更加关注对小节 bar 的生成而非单纯对 note 进行生成。作者从 Lakh MIDI dataset (LMD)数据集中挑选出了 10 万个 bar 来进行训练，生成 5 个轨道的琴键数据(piano-rolls)，分别是：贝斯,鼓,吉他，钢琴和弦乐。在预处理过程中，作者将钢琴琴键值转换为矩阵 X，每个 X 的矩阵大小是固定的，有利于 CNN 提取特征。同时还提出了一些量化标准来衡量生成的音乐质量。在训练过程中，根据 WGAN 的优化训练理论[28]，作者控制了 G 和 D 的更新次数，使得每跟新一次 G 则更新 5 次 D，以提升 D 的识别能力更好地训练 G。

MuseGAN 没有将和弦的先验条件考虑在内，这点与 MidiNet 有所不同。

3) GANsynth

文献[38]提出与图片数据不同，语音和音乐具有周期性的性质，如何保持周期信号的规律性，对生成结果的质量来说至关重要。作者提出在频谱域(spectral domain)对音乐数据进行对抗训练处理的模型 GANynth，相对于按序列生成音频的自回归模型，如 WaveNet，整体生成速度快了接近 5 万倍。

GANynth 主体使用了 Progressive GAN 的架构[24]，G 通过一系列卷积层将球型高斯分布中采样的噪声 z 样本数据逐步上采样，生成完整的声音张量后送入 D(架构为 G 的镜像)，由 D 判别其与真实分布的差异性。模型中采用 WGAN-GP 的梯度惩罚[28]来提升 Lipschitz 连续性。并且由于数据中增加了标签向量信息，因此也在 D 的损失函数中加入了 ACGAN[39]的辅助分类损失函数 auxiliary classification 来预测其标签。

相比于之前的工作 GANsynth 最大的创新点在于用一系列的频谱表示数据来训练 GAN。作者认为对于音乐这样体现出强烈周期性质的数据，为相位分量生成瞬时频率 (IF)的 GAN 优于其他的数据表示(如波形)，因此，GANsynth 并不

直接通过对抗生成波形文件，而是将生成的张量数据视为梅尔谱中间数据，再将该中间数据经过短时傅里叶变换转换为频谱数据，最后再根据频谱数据合成波形数据。

作者认为给定声波波形有自己的周期，波形一旦被反卷积或者短时傅里叶变换(STFT)等基于帧的技术处理，则又会有相应的帧周期概念，波形固有周期和帧周期之间无法精确对齐。对于反卷积来说很难覆盖到周期内所需的频率数据，也很难保证帧与帧之间的连贯性。如果采用 STFT 将波形转换为在频域中处理数据，则可以以 2π为周期来编码数据，并采样其顺势径向频率(instantaneous radial frequency)，将波形频率与帧频率很好地结合在了一起。

GANynth 模型采用的数据集是基于 NSynth dataset 的预处理样本,包含了来自 1000 种乐器的 300000 个乐器声音数据，该数据集具有大量的音色和音符数据，并包含了对这些数据的高度结构化标注信息，形成了音符、强度、乐器、听觉质量等丰富的标签。CelebFaces Attribute(简称 celebA)[45]在人脸属性相关任务领域中被大量使用，该数据集包含 200K 以上的名人脸部头像，每张样本都标注有 40 个属性，并且所有图片进行了裁剪对齐，保证数据分布的统一性。NSynth 数据集的制作动机是希望形成音乐数据集中的 celebA，聚焦于从乐器中提取出来的单个音符，从时间尺度、方差上进行归一化对齐(aligned)，使带训练模型更关注于数据本身特征(如音色，音准等)，同时对每个音符样本进行了 14 个类特征的标注，方便进行条件控制训练[41]。数据集中每个样本的时长被规范化为 4 秒，采样频率 16 kHz，维度为 64000，音调范围覆盖 MIDI 标准编码的 24~84。

实验结果显示，GANynth 模型对声谱图进行高频率分辨率采样的设置后(声谱图分辨率为[(128, 1024, 2)])，生成的音乐数据在人耳听觉，分布相似度距离，Inception 得分(借助 Inception 网络[44]进行判定的分数)等几个指标上面相对于 WaveNet 都有较大的优势。该项工作奠定了基于 GANs 对音频进行领域迁移(domain transfer)学习的基础，为包括语音在内的其他音频生成工作提供了借鉴意义。但是生成结果也表明，GANs 自身存在的一些问题在 GANynth 模型中依然没得到有效地解决，如模式塌陷。

2.3　结合强化学习的 GAN 生成音乐

正如文献[48]中所描述：“作曲系统可以朝着集多种方法为一体的混合型系统(hybridsystem)的方向发展”,将 GANs 作为一种训练框架，与其他机器学习或深度学习的方法联合，也是一种重要的研究方法。强化学习作为一种序列决策方法，其训练过程可以描述为通过动态调整自身状态采取行动以获取奖励的过程。近年来，得益于大数据的普及、计算力的提升及新算法演进，特别是与深度学习的结合，使强化学习在游戏博弈，机器翻译，文本序列预测等领域取得了一定突破。同时在将强化学习与生成对抗机制结合，使二者特性互补的方面，也有工作作出了尝试，如 2.4 节所述的 SeqGAN 就是其中的典型代表。SeqGAN 解决了离散数据在对抗训练过程中不容易回传梯度更新生成器的问题，并借助蒙特卡洛搜索来补全瞬时生成的序列数据，在文字生成和音乐生成等实际应用上也给出相应的范例。文献[2]在 SeqGAN 的基础上，更进一步扩展了其在音乐创作层面的应用范畴，对生成复调音乐的方式做了相应的探索(以下将该工作的模型称为 Polyphonic SeqGAN)。

Polyphonic SeqGAN 认为复调音乐的生成，特别是基于和弦约束的复调音乐生成,能更好地提升音乐整体生成质量。作者用一种高效的数据组织方式动态获取乐曲的旋律和和弦，将音符，音符时长，和弦等数据信息封装到 word vector 中，训练模型在音乐词嵌入空间(the embedded musical word space)中预测旋律序列的分布。G 模型的主干网络采用 RNN，D 模型的主干网络采用 CNN，同时将原始 GAN 中的交叉熵损失函数替换为最小二乘 GANs[42]中提出的最小二乘损失，以提高模型的稳定性和收敛性。G 在每一个时间步(time step)的生成结果序列都会被补全后交由 D 进行评判，由 D 给出得分作为奖励信号，G 基于该奖励信号进行更新优化。

为了增强对复音音乐的生成能力，作者借鉴了 performanceRNN(https://magenta.tensorflow.org/performance-rnn)的对音乐数据的表达方式，训练数据与 SeqGAN 采用了相同的数据集 Nottingham，但与之不同的是，SeqGAN 从数据集中挑选出的是固定时间步长的单调音乐样本，PolyphonicSeqGAN 则使用复调样本来进行训练。作者从 midi 样本中抽取的元数据被分为和弦和音符两个种类，音符数据从起始时间，时长，音符值等维度进行记录，和弦也由组成的根音、三音等元素进行记录形成多维向量组，随后通过词典(Vocabulary)映射的方式转换为嵌入向量，送入模型。

PolyphonicSeqGAN 的生成结果显示虽然生成对抗机制虽然在很大程度上提升了音乐空间的建模处理能力，但是 GANs 中一些天生缺陷，如模式崩塌造成生成的乐曲单一化(G 采用覆盖小范围分布的生成方式来欺骗 D 而不是尽力拟合真实数据分布)，作者也提出可以采用 WGAN 的 EM 距离来修正该缺陷[27][28]。同时，由于蒙特卡洛树搜索的随机性质，使得生成结果变化较大且具有不可重复性。另外，模型生成效率不高,相对于负对数似然估计(log-likelihood NLL)生成来说，要慢十倍左右。

3　基于生成对抗网络进行音乐生成技术的现状及展望

GANs 为研究者提供了一种对抗训练的框架和思想，而没有具体限制模型实现的方法，因此具备较强的灵活性和扩展性,可以将任何合适的主干网络和损失函数融入该框架中。众多音乐生成的工作充分利用了 GANs 的优势和特点，在原生 GANs 的基础上，将 RNN，CNN，强化学习 RL 等模型扩展应用至对抗框架中，取得了较好的效果。据观察：

a) 采用对抗训练的模型生成音乐效果明显优于非对抗训练的生成效果[33, 36]；

b) 对音乐生成的最小关注单位从音符(note)向小节(bar)过度[5, 1]，基于后者进行生成的模型在质量上往往好于前者，这也是乐理约束的体现之一；

c) 原生 GANs 具有的一些先天不足也会传递到对应的音乐制作模型中，最显著的一点即为模式塌陷(mode collapse)，如文献[2, 29]。同时，随着 GANs 技术本身的提升，也相继出现一些针对这些缺陷的解决或缓解方案，并应用到相应的音乐制作工作中[1]；

d) 对于数据集的关注度越来越大，这也是算法模型发展的必然结果，除了早期的 Nottingham 数据集外，一些更为精心设计的音乐样本标注数据集也得了发展和应用，如包含了 30 万个样本的 NSynth 数据集，其构建目的是为了达到 celebA 数据集在人脸领域中的精细化属性标注数据集的地位。

对于今后的研究方向，可能存在以下几个趋势：

a) 随着生成对抗网络技术的发展，最新的研究成果将会越来越快越来越广地适配到音乐制作领域中，已达到弥补 GANs 的部分先天缺陷及提升生成质量的目的，如更鲁棒的损失函数，模式崩塌的缓解，框架中结合更前沿的学习算法等；

b) 和乐理先验知识的结合将会越来越紧密，特别是其中关联性较大的元素，如和弦，节奏，强弱拍。经过对乐理知识的梳理后认为，在影响音乐生成质量的众多乐理因素中，

和弦走向及和弦匹配度是至关重要的，和弦走向奠定了整首歌曲旋律的框架，目前尚未有模型对其做单独的精细化生成，引入对抗训练机制后，将有助于提升和弦走向的生成同时也促进后续音符的生成质量提升。基于此，可尝试提出一个新的模型，在 GANs 框架下由粗(对抗生成和弦走向)至细(以和弦走向为隐形约束对抗生成旋律)分阶段地对音乐数据进行生成；

c) 当前引入对抗机制进行音乐生成的方法大多属于随机生成或‘弱控制’生成的范畴，部分模型加入了和弦等控制因素，但整体上来说在参数可解释性上还是较弱，后期可借鉴生成对抗模型特别是人脸生成领域中的一些精细化控制生成方法[43]，在生成参数的可解释性上做进一步研究。

d) 算法模型的技术更新对音乐标注数据集的发展也起到了一定需求推动作用，音乐数据集将朝着精细化，规模化的方向快速发展，并更好地促进算法模型的迭代更新，形成良性循环。

参考文献：

[1] Dong H W, Hsiao W Y, Yang L C, *et al.* MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment [J]. 2017.

[2] Lee S G, Hwang U, Min S, *et al.* Polyphonic Music Generation with Sequence Generative Adversarial Networks [J]. 2017.

[3] Chu H, Urtasun R, Fidler S. Song From PI: A Musically Plausible Network for Pop Music Generation [J]. 2016.

[4] Oord A V D, Dieleman S, Zen H, *et al.* WaveNet: A Generative Model for Raw Audio [J]. 2016.

[5] Yang L C, Chou S Y, Yang Y H. MidiNet: A convolutional generative adversarial network for symbolic-domain music generation [J]. 2017.

[6] Alpern, Adam (1995) , Techniques for algorithmic composition of music. http://alum. hampshire. edu/~adaF92/algocomp/algocomp95. html. Hampshire College.

[7] Douglas Eck and Juergen Schmidhuber. A first look at music composition using lstm recurrent neural networks. 2002.

[8] Fernandez J D, Vico F. AI Methods in Algorithmic Composition: A Comprehensive Survey [J]. Journal of Artificial Intelligence Research, 2014, 48 (1) .

[9] Hiller L, Isaacson L. Musical composition with a high-speed digital computer [M]// Machine models of music. MIT Press, 1992.

[10] Chan, Michael & Potter, John & Schubert, Emery. (0002) . Improving algorithmic music composition with machine learning.

[11] Peter, M, Todd. A Connectionist Approach to Algorithmic Composition [J]. Computer Music Journal, 1989.

[12] Todd P M B J J. Modeling the perception of tonal structure with neural nets [J]. Computer Music Journal, 1989, 14 (4): 44-53.

[13] MOZER, Michael C. Neural Network Music Composition by Prediction: Exploring the Benefits of Psychoacoustic Constraints and Multi-scale Processing [J]. Connection Science, 1994, 6 (2-3): 247-280.

[14] Chen CCJ; Miikkulainen R. Creating melodies with evolving recurrent neural networks [Z]. , 2001.

[15] Kang S, Ok S Y, Kang Y M. Automatic Music Generation and Machine Learning Based Evaluation [M]// Multimedia and Signal Processing. Springer Berlin Heidelberg, 2012.

[16] Allen Huang and Raymond Wu. Deep learning for music. arXiv preprint arXiv: 1606. 04930, 2016

[17] Douglas Eck and Juergen Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on, pages 747–756. IEEE, 2002.

[18] Pascal Vincent Nicolas Boulanger-Lewandowski, Yoshua Bengio. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In Proceedings of the 29th International Conference on Machine Learning (ICML) , page 1159–1166, 2012.

[19] Goodfellow I, Pougetabadie J, Mirza M, *et al.* Generative Adversarial Nets [C]. neural information processing systems, 2014: 2672-2680.

[20] 邹秀芳，朱定局. 生成对抗网络研究综述. 计算机系统应用，2019, 28 (11): 1-9 (ZOU Xiu-Fang, ZHU Ding-Ju. Review on Generative Adversarial Network. COMPUTER SYSTEMS APPLICATIONS, 2019, 28 (11): 1-9)

[21] Anders T, Miranda E. Constraint programming systems for modeling music theories and composition [J]. ACM Computing Surveys, 2011, 43 (4) .

[22] Pearce M T, Meredith D, Wiggins G A, *et al.* Motivations and Methodologies for Automation of the Compositional Process [J]. Musicae Scientiae, 2002, 6 (2): 119-147.

[23] Radford A, Metz L, Chintala S, *et al.* Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks [C]. international conference on learning representations, 2016.

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In ICLR, 2018a.

[25] Liu Z, Luo P, Wang X, *et al.* Deep Learning Face Attributes in the Wild [C]. international conference on computer vision, 2015: 3730-3738.

[26] Arjovsky M, Léon Bottou. Towards Principled Methods for Training Generative Adversarial Networks [J]. Stat, 2017, 1050.

[27] Arjovsky M, Chintala S, Bottou L, *et al.* Wasserstein Generative Adversarial Networks [C]. international conference on machine learning, 2017: 214-223.

[28] Gulrajani I, Ahmed F, Arjovsky M, *et al.* Improved Training of Wasserstein GANs [J]. arXiv: Learning, 2017.

[29] Yu L, Zhang W, Wang J, *et al.* SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient [J]. 2016.

[30] Manaris, B. ; Roos, P. ; Machado, P. ; *et al.* 2007. A corpus-based hybrid approach to music analysis and composition. In NCAI, volume 22, 839 Deep Artificial Composer: A Creative Neural Network Model for Automated Melody Generation

[31] Douglas Eck and Juergen Schmidhuber. Finding temporal structure in music: Blues improvisation with lstm recurrent networks. In Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on, pages 747–756. IEEE, 2002.

[32] Pascal Vincent Nicolas Boulanger-Lewandowski, Yoshua Bengio. Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription. In Proceedings of the 29th International Conference on Machine Learning (ICML) , page 1159–1166, 2012.

[33] Mogren O. C-RNN-GAN: Continuous recurrent neural networks with adversarial training. [J]. arXiv: Artificial Intelligence, 2016.

[34] Emily L Denton, Soumith Chintala, Rob Fergus, *et al.* Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in neural information processing systems, pages 1486–1494, 2015.

[35] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In Proc. Advances in Neural Information Processing Systems, pages 2226–2234, 2016.

[36] Trieu, Nicholas & Keller, Robert. (2018) . JazzGAN: Improving with Generative Adversarial Networks.

[37] Hochreiter, Sepp, Schmidhuber, Jürgen. Long Short-Term Memory [J]. Neural Computation, 9 (8): 1735-1780.

[38] Engel J, Agrawal K K, Chen S, *et al.* GANSynth: Adversarial Neural Audio Synthesis [J]. 2019.

[39] Odena A, Olah C, Shlens J, *et al.* Conditional Image Synthesis With Auxiliary Classifier GANs [J]. arXiv: Machine Learning, 2016.

[40] Eitan Richardson and Yair Weiss. On GANs and GMMs. CoRR, abs/1805. 12462, 2018. URL http://arxiv. org/abs/1805. 12462.

[41] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with WaveNet autoencoders. In ICML, 2017

[42] Mao X, Li Q, Xie H, *et al.* Least Squares Generative Adversarial Networks [J]. arXiv: Computer Vision and Pattern Recognition, 2016.

[43] Ma D, Liu B, Kang Z, *et al.* Two birds with one stone: Transforming and generating facial images with iterative GAN [J]. Neurocomputing, 2019.

[44] Szegedy C, Liu W, Jia Y, *et al.* Going deeper with convolutions [C]. computer vision and pattern recognition, 2015: 1-9.

[45] Liu Z, Luo P, Wang X, *et al.* Deep Learning Face Attributes in the Wild [C]. international conference on computer vision, 2015: 3730-3738.

[46] 王程，周婉，何军. 面向自动音乐生成的深度递归神经网络方法 [J]. 小型微型计算机系统，2017 (10) . (Wang cheng, Zhou wan, He jun. Method of auto-music-generation oriented recursive neural network [J]. J Chin Mini-Micro Comput Syst. 2017 (10)

[47] 苗北辰，郭为安，汪镭. 隐式特征和循环神经网络的多声部音乐生成系统 [J]. 智能系统学报，2019, 14 (01): 162-168. (Miao beechen, Guo weian, Wang lei. Polyphonic music generation system of latent features and recurrent neural network [J]. CAAI T Intell Syst, 2019, 14 (01): 162-168)

[48] 冯寅，周昌乐. 算法作曲的研究进展 [J]. 软件学报，2006, 17 (2): 209-215. (Feng yin, Zhou change. Research progress of algorithmic composition [J]. Journal of Software, 2006, 17 (2): 209-215.)

[49] 朱纯，王翰林，魏天远等. 基于深度卷积生成对抗网络的语音生成技术% 仪表技术，2018, 000 (002): 13-15, 20. (Zhu chun, Wang hanlin, Wei tianyuan, *et al.* Speech Generation Based on Depth Convolution for Adversarial Networks [J]. Motormeter Technology, 2018, 000 (002): 13-15, 20)

[50] Zhao Z, Sun Q, Yang H, *et al.* Compression artifacts reduction by improved generative adversarial networks [J]. EURASIP Journal on Image and Video Processing, 2019, 2019 (1): 1-7.

[51] Zhang H, Xu T, Li H, *et al.* StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks [C]. international conference on computer vision, 2017: 5908-5916.

[52] Bulat A, Yang J, Tzimiropoulos G, *et al.* To Learn Image Super-Resolution, Use a GAN to Learn How to Do Image Degradation First [C]. european conference on computer vision, 2018: 187-202.